New Challenges in Environmental Statistics

Nuove sfide per le statistiche ambientali

Peter Guttorp
University of Washington, Seattle, USA

Riassunto: Quello delle statistiche ambientali è un campo in rapido sviluppo. Il presente lavoro apporta alcuni esempi dei principali problemi che possono sorgere e dei metodi utilizzati per analizzare i dati nel contesto della scienza ambientale.
I dati ambientali spesso derivano dal monitoraggio dello spazio e del tempo. Di conseguenza, l'analisi di questi dati richiede l'uso di strumenti che tengano conto sia della dipendenza spaziale che di quella temporale. In questo lavoro vengono, quindi, mostrati alcuni di questi strumenti di analisi spazio-temporale, recentemente sviluppati per lo studio dei dati sull'inquinamento atmosferico e sul clima in generale.

Keywords: environmental statistics, space-time models

1. Introduction

The field of environmental statistics is one of rapid growth. There are scientific organizations (both the American Statistical Association and the Royal Statistical Society have sections dealing with statistics and the environment, and there is an international society called TIES, the International Environmetric Society), as well as conferences and scientific journals. The field is fairly broad in terms of methodology, and in this presentation I will give a few examples of problems and methods for analyzing data in the context of environmental science.

Much recent work in the field deals with measurements that are expressed both in time and space. In section 2 I describe some recent work at the University of Washington (done by my colleague Paul Sampson) on estimating space-time fields of air quality from monitoring data. The topic of analyzing space-time extremes in climate is briefly discussed in section 3. The hierarchic approach to state-space modeling has been very successful in dealing with geophysical models. It can also be used to deal with the problem of combining measurements made at different spatial and temporal scales, as is illustrated in section 4 by the work of Tamre Cardoso, a recent PhD student of mine, on modeling precipitation rates based on different types of measurements. Finally, in section 5 I sketch how one can think about the statistical quality of environmental standards based on health effects analysis of air quality data.

2. Space-time models

Environmental data often arise from monitoring in space and time. Consequently, the analysis of these data requires tools that can deal with both spatial and temporal dependence. In this section we show some spatio-temporal tools that have recently been developed for analysis of air pollution data.

In geophysics and meteorology, variants of principal components called empirical orthogonal functions (EOFs) have long been used to describe leading modes of variability in space-time processes (e.g., North, 1984). Here we use smoothed EOFs to model the spatio-temporal mean of a random field viewed as spatially varying systematic temporal trends. It is common in the space-time modeling literature to decompose observations into the sum of a systematic trend component and residuals

$$Z(x,t) = \mu(x,t) + \varepsilon(x,t) .$$

Where details of the trend structure vary spatially, we write such a decomposition more generally, with spatially varying coefficients, as

$$\mu(x,t) = \beta_{x0} + \sum_{j=1}^{J} \beta_{xj} f_j(t) ,$$

the $\{f_j(t)\}$ being a set of orthogonal temporal basis functions. Our focus is on applications with smooth seasonal trends. Many air quality parameters display a dominant seasonal trend structure that is not conveniently represented by sums of trigonometric basis functions. In these cases we seek a parsimonious set of nonparametric basis functions $\{f_j(t)\}$, where the first basis function $f_1(t)$ typically represents the dominant or average trend over the spatial region of interest, and subsequent basis functions, computed to be orthogonal to the first, along with the spatially indexed parameters β_{xj} permit the shape and amplitude of the spatial structure to vary. Technically, the basis functions are computed by running a nonparametric smoother along the singular value decomposition of the data matrix (which has been completed by an EM-like algorithm; see Guttorp et al., 2005, for details).

The analysis of 8-hour maximum average daily ozone concentrations from southern California for the period 1987-94 was one of seven similar analyses of data from regions spanning most of the continental United States. Ozone seasonal trends are similar nationwide and we hoped to be able to use a single set of temporal trend basis functions for all regions. The computation of trend components was based on data from 513 monitoring sites monitoring nearly throughout the year across the country. Analysis of the ozone concentration data was judged most appropriate on a square root scale. Figure 1 shows the four temporal basis functions computed from the first four left singular vectors of the 2912 × 513 data matrix. The first function clearly represents the dominant seasonal ozone cycle with highest concentrations during the sunny summer months. The shape and amplitude of this seasonal feature varies from year to year, another notable distinction with many seasonal trend models, which do not adapt to fluctuations in trend. The second trend component is necessary for many sites, which display a pair of ozone peaks, one in spring (due to influx of ozone from the stratosphere) and another later in the summer (due to photochemical creation of ozone in the troposphere). The third and fourth components serve mainly to adjust the exact shape and locations of the seasonal peaks defined by the first two components.

Approximately 67% of the variance in the entire (scaled) 2912 × 513 data matrix is explained by the first four unsmoothed components.

Figure 1: *First four singular vectors (dots) and smooth trend components derived from the 2912 x 513 matrix of square root transformed ozone.*

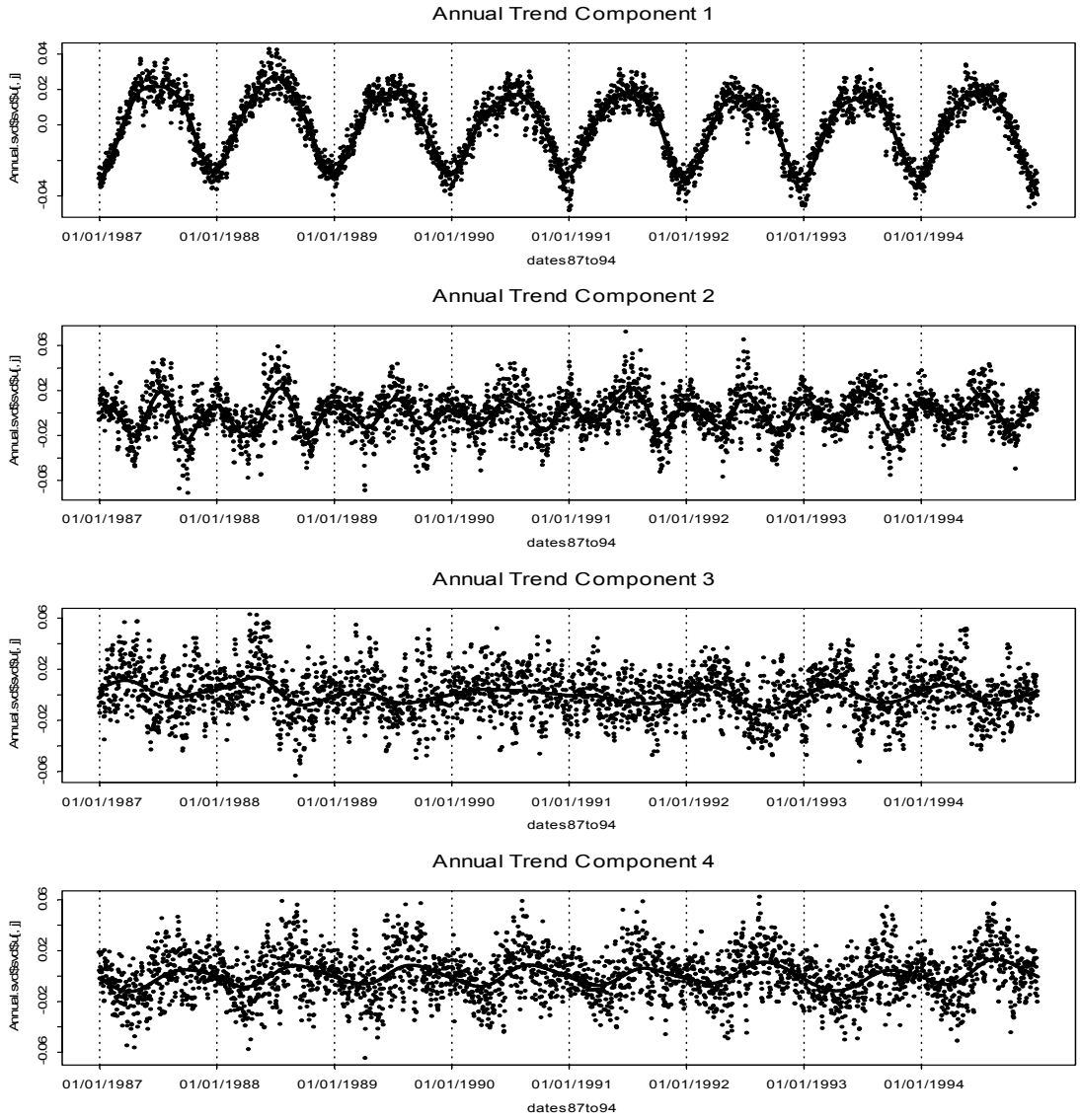
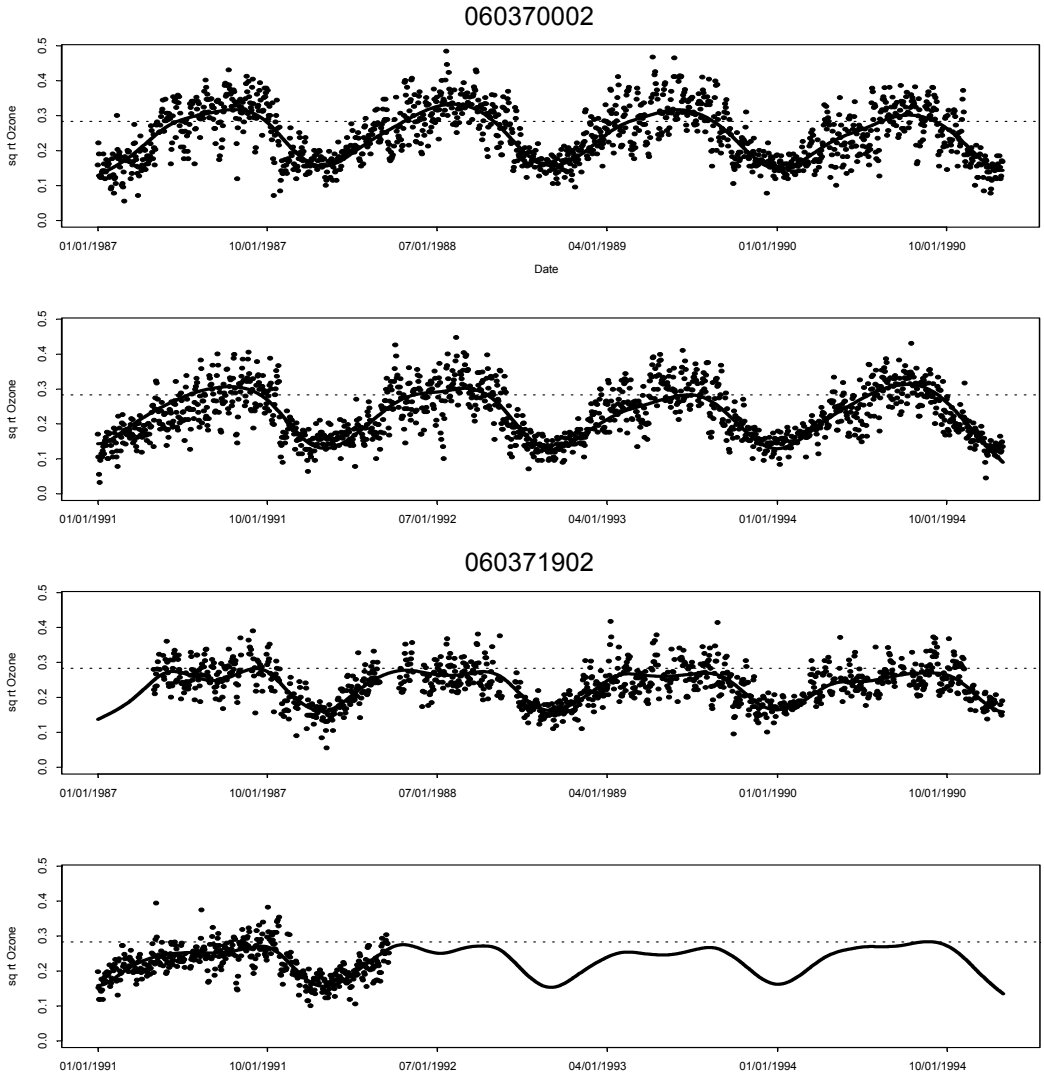


Figure 2 illustrates trends for two monitoring sites in Los Angeles County. These were computed as linear combinations of the four trend components illustrated in Figure 1. We note that there is distinct variation in the shape of the seasonal trend even over this relatively small spatial region with the first site, 060370002, showing the dominant seasonal pattern and the second site, 060371902 (which was inoperative in early 1987 and from 1992 through 1994), showing two reasonably distinct seasonal maxima.

Figure 2: *Fitted temporal trends for two monitoring sites in Los Angeles county.*



The coastline, complex topology, and typical weather patterns combine to effect a complex nonstationary spatial correlation structure in the spatio-temporal residuals from the fitted temporal trends. Sampson and Guttorp (1994) introduced an approach to nonstationary spatial covariance modeling in which the geographic coordinates are deformed to create a geography (the dispersion plane, or D-plane) in which the covariance structure is approximately isotropic. This approach is usually applied to detrended residuals. We assume for simplicity that the temporal structure of the residuals

$$\hat{\varepsilon}(x,t) = Z(x,t) - \hat{\mu}(x,t)$$

is white noise. We decompose the residuals

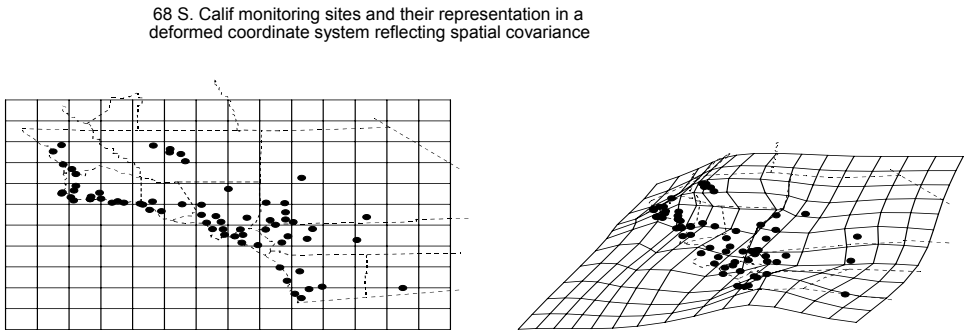
$$\delta(x,t) = \nu(x)\, H(x,t) + E(x,t),$$

where $H(x,t)$ is a mean zero, variance one spatial process with covariance structure

$$\mathrm{Cov}(H(x,t), H(y,t)) = \rho_\theta\left(\left| f(x) - f(y)\right|\right)$$

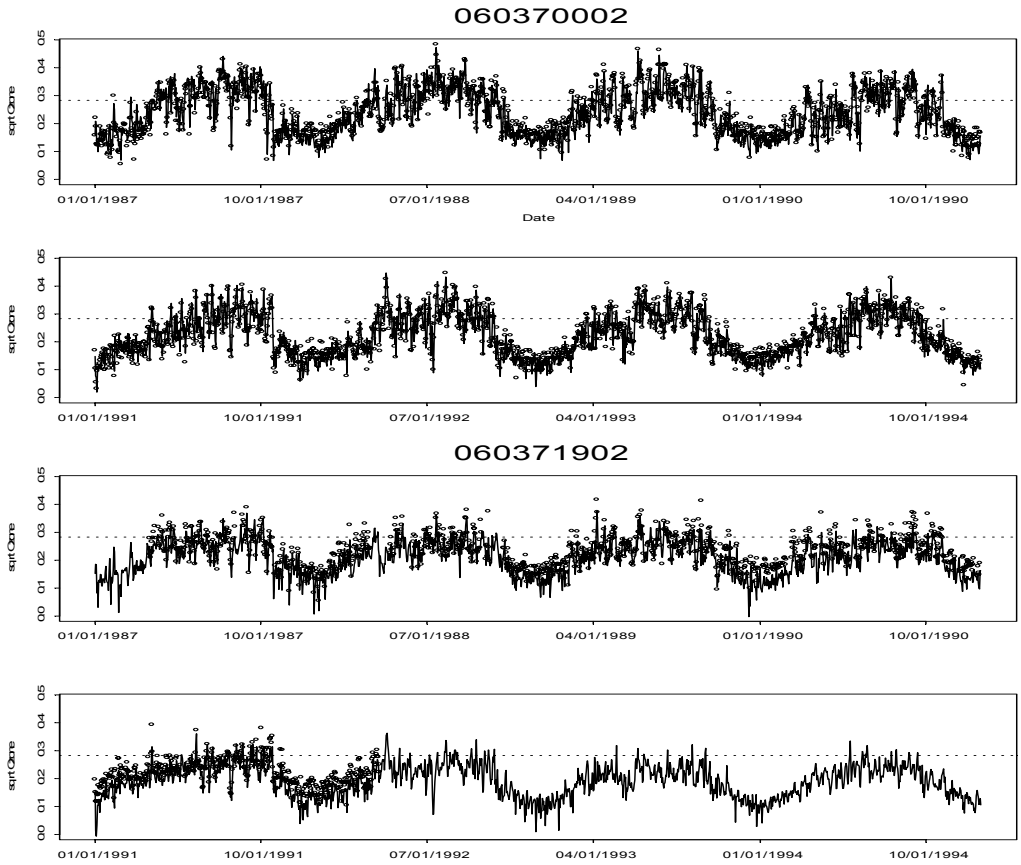
and $E(x,t)$ is a white noise process, uncorrelated with $H(x,t)$. The function f is the deformation of the geographic plane, and is fitted using a pair of thin-plate splines (Bookstein, 1989). Technically we use a Gaussian-based Bayesian approach using MCMC, detailed in Damian et al. (2000,2003). This has the advantage that we can draw samples from the deformations, and get a good feeling for the uncertainty in the fit. Figure 3 depicts the posterior mean estimate of the spatial deformation computed to permit fitting of stationary isotropic correlation models in the deformed coordinate system. The predominant feature of compression along the coastline running NW-SE indicates that spatial correlation is strongest parallel to the coast and weaker orthogonal to the coast.

Figure 3: *Posterior mean spatial deformation representing nonstationary spatial covariance structure.*



Spatial prediction (kriging) of the trend coefficients combined with prediction of the spatio-temporal residuals produce the cross-validation predictions of the time series for the two LA county sites, as illustrated in Figure 4.

Figure 4: *Cross-validation predictions of daily (square root) ozone concentrations at two monitoring sites in Los Angeles county.*



3. Extreme events

There is a lot of current interest in judging whether we are observing changes in the Earth's climate due to global warming. A special issue of the *Bulletin of the American Meteorological Society* (no. 3, 2000) contains five articles about trends in extreme weather and climate events, covering observations (Easterling et al., 2000), socioeconomic impacts, terrestrial ecological impacts, and model predictions (Meehl et al., 2000). The tools used in these articles are those developed for single variable extremes in independent and identically distributed data, and lack firm estimates of variability. There is a rapidly growing amount of data on extremes, both from observations and particularly from climate models, which are now being used in ensemble predictions of future climate (e.g. Stainforth et al. 2005). However, the statistical methods used to interpret this information including trend assessments of extreme events, requires some attention. Several gaps in the knowledge were identified and future research needs pointed to in IPCC (2002). The use of spatial data can help to improve estimation of extreme value models that are regionally similar (Casson and Coles, 1999).

Classical multivariate extreme-value theory is based on asymptotic theory, which is valid when looking at extremes in all components simultaneously. This is often not appropriate for the applications to climate extremes. Rather, the conditional approach introduced by Heffernan and Tawn (2004) can be used to compute probabilities of various extreme events. Here one looks at the conditional distribution of the other variables, given that one is extreme. Events of interest may include the probability of obtaining events in fairly general sets in multivariate space. The sets of interest can, e.g., be developed by using various index numbers that describe combinations of atmospheric variables that induce physical stress on some part of the environment (e.g. Frich et al., 2002). For example, determination of drought conditions can be done using a drought index (Heim 2002 is an overview), most of which are linear combinations of variables relating to the hydrologic cycle, such as precipitation, soil moisture at different depths, and evapotranspiration. Hence extreme values for a drought index correspond to a complex set in multivariate space. It is then of interest to look at issues such as the probability of severe drought conditions computed from a climate run with changing forcing of the climate system. Due to seasonal components of the hydrologic cycle, one needs to consider seasonally changing extreme value distributions, and thus seasonally varying sets corresponding to severe drought. Tools developed for univariate seasonally varying extremes must then be extended to the multivariate case.

4. The state-space approach

In many situation a statistician would naturally think of the 'true state of nature' as a parameter we are interested in estimating. Often there are different measurement instruments, yielding data collected on different spatial and temporal scales. The combination of such data into a single analysis can often be done using state-space models. We describe a general setup for this hierarchical way of setting up models. Then we discuss how one can combine different measurements of precipitation into an automatically calibrated "ground truth" estimate.

4.1 A general hierarchical setup

Consider an unobserved "real" entity X. If we knew X we would often be able to write down a conditional distribution of observations Y, given X and some parameter θ. Letting $[Z]$ stand for the distribution of Z, we can write this hierarchical model in shorthand as

$$[Y|X,\theta] \quad \text{Observations}$$
$$[X|\theta] \quad \text{State equation}$$
$$[\theta] \quad \text{Parameters}$$

In typical geophysical models, the state equation is described using a system of partial differential equations, but frequently it can be approximated successfully by a relatively simple stochastic model, such as a multivariate autoregressive model (see Berliner, 2000).

4.2 Estimating precipitation

There are several different ways to measure precipitation. The most common is to use rain gauges. These measure the amount of rain over usually a moderate amount of time, such as 24 hours, and are very local in their spatial coverage, i.e., if we put another gauge right next to the first one if will measure a different amount. There are know sources of bias for gauges, such as underestimating low rainfall rates due to evaporation, and underestimating amounts in high winds. Another tool for measuring precipitation is radar. The radar operates by measuring the reflectivity in the air, and can cover large spatial areas every few minutes. A third ground-based instrument is the distrometer, which measures the drop size distribution. This instrument is as local as the rain gauge, and has higher temporal resolution than any other instrument. There are also a variety of satellite measurements of precipitation. The key scientific question is how to calibrate these instruments to obtain the most accurate measurement of rain rate.

Accurate measurements of precipitation are needed for input to hydrological models, for agricultural planning, and for various other engineering purposes. Theoretically, all quantities of interest (such as the rain rate $R(t)$ and the reflectivity $Z_D(t)$) are functions of the drop size distribution. We can write

$$R(t) = c_R \frac{\pi}{6} \int_0^\infty D^3 v(D) N(t) f(D) dD$$

and

$$Z_D(t) = c_Z \int_0^\infty D^6 v(D) N(t) f(D) dD$$

where v is the terminal velocity of a rain drop, N is the number of drops, and f is the probability density of drop sizes. In other words, the drop size distribution is described as the total number of drops and the density of drop size, given the total number of drops. Observational data indicate that the density of drop size can reasonably be taken as independent of the total number of drops. If we have distrometer data, it is then reasonable to use them as “ground truth” and calibrate the other instruments relative to the distrometer. However, distrometers are expensive instruments, and not generally available for routine measurements. From a statistical point of view this is not a problem: we simply think of the drop size distribution and the total number of drops as state variables, and estimate them from radar and gauge data. The observation model is then

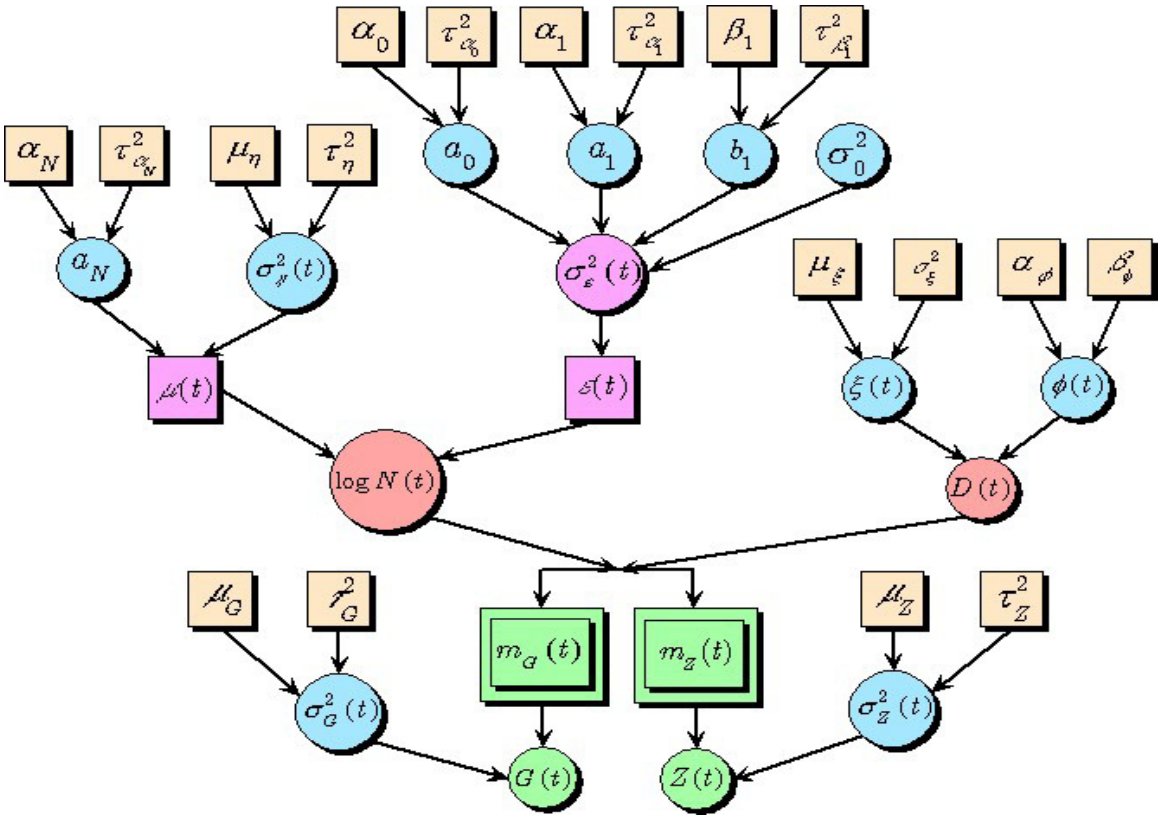
$$[G(t) \mid N(t), D(t)] \sim N\left(g(w(t)) \int_{t-\Delta}^{t} R(s) ds, \sigma_G^2 \right)$$

where g is a bias correction factor for gauges and w is wind speed (and possibly other covariates), and

$$[Z(t) \mid N(t), D(t)] \sim N(Z_D(t), \sigma_Z^2).$$

The state equation is described in three steps. The total number of drops follows a GARCH-model, with mean that is described by an AR(1)-model. Conditionally upon the total number of drops, the drop size density if given by a lognormal distribution. Prior distributions for all the parameters are then specified in order to fit the model to data using Markov chain Monte Carlo. The structure of the model is given in Figure 5.

Figure 5: *Structure of MCMC estimation for precipitation.*



The model was applied (Cardoso, 2004) to data from Eureka, California. A wintertime precipitation event of 550 minutes was modeled using this approach. Figure 6 shows the model estimate of $\log_{10}$ rain rate (not using the distrometer data but only gauge and radar) with observed values calculated from the distrometer data shown as dots

Figure 6: *Model estimates (histogram) and observed (dots) rain rates for Eureka, California.*

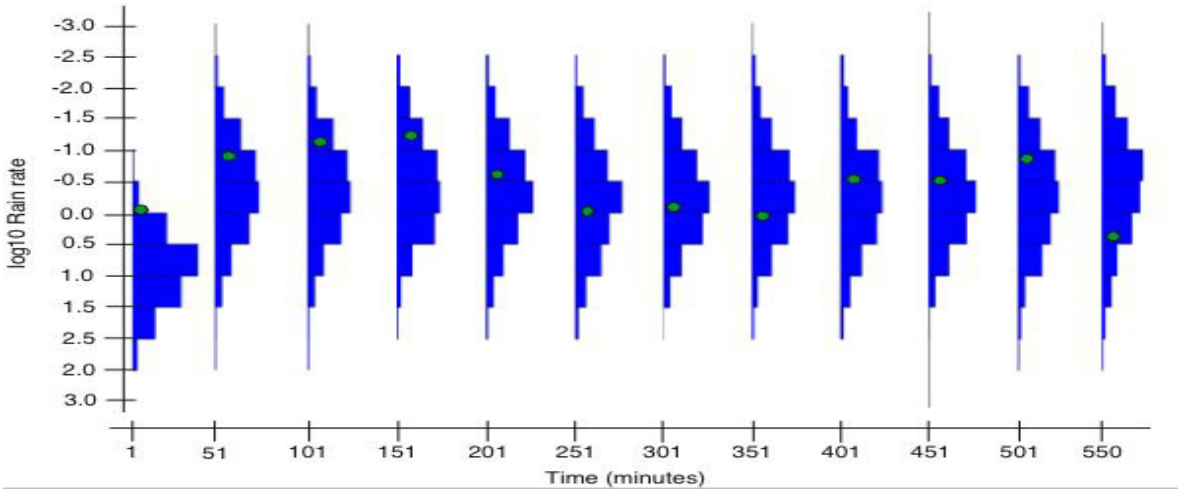
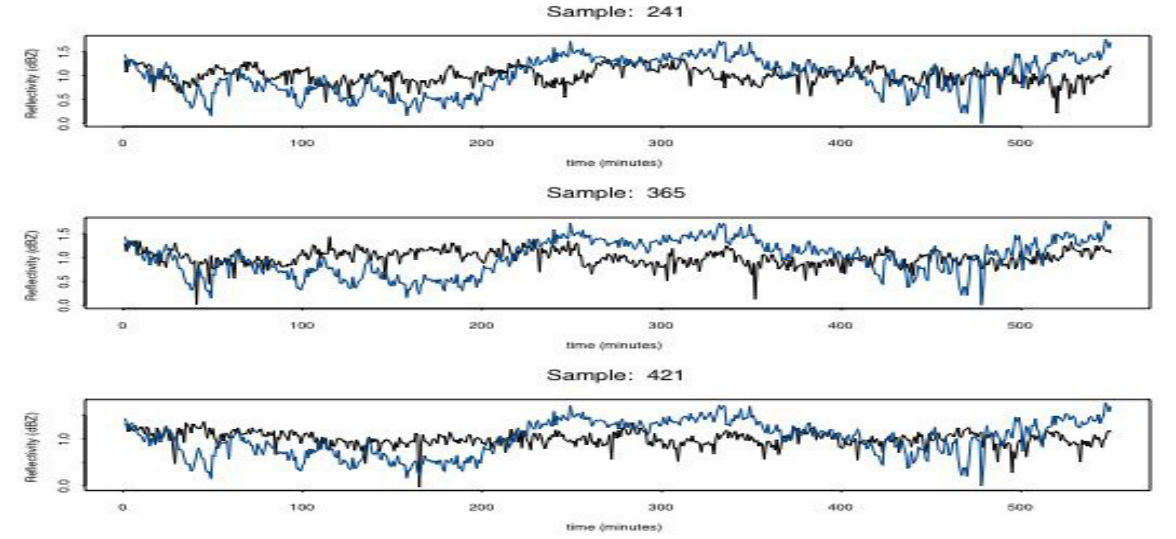


Figure 7 shows the estimated radar reflectivity from the model, together with the computed reflectivity from the distrometer data.

Figure 7: *Normalized (to mean 1) posterior mean reflectivity for selected samples. The blue lines show estimated reflectivity from distrometer data.*



5. Setting environmental standards

Most air quality standards are set by deciding upon some limit value, beyond which damage is thought to be done to the population, or to sensitive subpopulations. Such limit values are generally derived from health effect studies. Typically, they are implemented by requiring that measurements shall be below the limit value.

Given the limit value (which of course has some uncertainty associated with it), one can set up the issue of environmental standards as a statistical decision problem. In fact, many standards can be thought of from the point of view of a classical Neyman-Pearson test (cf. Cox et al., 1999). The null hypothesis, given that in most places the environmental agencies are charged first and foremost with protecting human health, must be that a region is in violation of the standard. There is often some leeway for the region, in that it is allowed to violate the standard only a limited amount of times. It then becomes possible to assess the implementation of a standard by statistical quality criteria, such as the probability of type I and type II errors (Barnett and O'Hagan, 1997).

However, if the implementation of a standard is in terms of measurements at a small number of monitoring stations, it is important to think about the conditional distribution of maximum exposure in a region, given these measurements. Again we are interested in extreme value computations, as in section 3. What is, for example, the conditional distribution of maximum exposure in a region for which the monitoring station is representative? Here we need to define representative, which we can do in terms of spatial correlation. Again, a space-time model, but now describing extreme values, is called for.

As an example, calculations by Sofia Åberg, a PhD student at the Lund School of Technology, indicate that for parameter values typical for Houston (Texas, USA) tropospheric hourly maximum ozone measurements, given that a measurement is just at the value 120 ppb which is the current limit for that site, and allowing the representative area to be one in which the correlation (at the square root scale) with the measurement site is at least 0.7, the conditional probability of being exposed to more than 120 ppb is about 0.63, and the conditional probability of being exposed to more than 180 ppb (at which serious health effects are expected) is about 0.05. These types of calculations are a different aspect of the statistical quality of the implementation of a standard, and one that so far has only received limited attention.

In addition, the health effect studies for air pollution are mostly opportunistic studies, using monitoring data from networks designed for detecting large values (i.e., standard violations) while the health data often result from cohort studies that are specifically designed for other health risks. There is a danger in this, since the health effects that can be specifically related to air quality generally are rather small, and a network designed to find large values (at least if it is successful at its purpose) would tend to overestimate the exposures of individuals not living at the monitoring site, and hence underestimating the health effects at a given level of exposure. The size of this type of design bias is presently unknown, and should be investigated carefully.

References

Barnett, V. and O'Hagan, A. (1997) *Setting environmental standards: the statistical approach to handling uncertainty and variation*. London: Chapman & Hall.

Berliner, M. (2000), Hierarchical Bayesian modeling in the environmental sciences. *Allgemeines Statistisches Archiv*, 84(2) , 141-153.

Bookstein, F. L. (1989) Principal warps–Thin-plate splines and the decompositions of deformations. IEEE Trans. *Pattern Analysis Mach. Intelligence* 11(6), 567–585.

Cardoso, T. (2004) *A hierarchical Bayes model for combining precipitation measurements from different sources*. Unpublished PhD dissertation, Graduate Program of Quantitative Ecology and Resource Management, University of Washington.

Casson, E. and Coles, S. (1999) Spatial Regression Models for Extremes. *Extremes* 1, 449-468.

L. H. Cox, P. Guttorp, P. D. Sampson, D. C. Caccia and M. L. Thompson (1999): Preliminary statistical examination of the uncertainty and variability on environmental regulatory standards for ozone (with discussion). In: *Environmental Statistics: Analyzing Data for Environmental Policy*. Novartis Foundation. Chichester: John Wiley & Sons, Ltd. 122-143.

D. Damian, P. D. Sampson and P. Guttorp (2000): Bayesian estimation of semi-parametric non-stationary spatial covariance structure. *Environmetrics* 12,161-176.

D. Damian, P. D. Sampson, and P. Guttorp (2003): Variance modeling for nonstationary spatial processes with temporal replication. *Journal of Geophysical Research Atmospheres* 108(D24) Art. No. 8778.

Easterling, D. R., Evans, J. L. Groisman, P. Ya., Karl, T. R. et al. (2000): Observed variability and trends in extreme climate events: A brief review. *Bull. Amer. Met. Soc*. 81(3), 417-426.

Frich, P., L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. M. G. Klein Tank, and T. Peterson (2002) Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Research*, 19, 193-212.

Guttorp, P., Fuentes, M. and Sampson, P. D. (2005) Using transforms to analyze space-time processes. Submitted to: *SEMSTAT Proceedings of a workshop on space-time analysis.* Also available as NRCSE TRS 80 at http://www.nrcse.washington.edu/pdf/trs80.pdf.

Heffernan, J. E. and Tawn, J. A. (2004) A conditional approach for multivariate extreme values (with Discussion). *J. Royal Statist. Soc*. Series B 66, 497-546.

Heim (2002) A Review of Twentieth-Century Drought Indices Used in the United States. *Bull. Amer. Met. Soc.* 83(8), 1149-1165.

IPCC (2002) *IPCC Workshop on changes in Extreme Weather and Climate Events*. Beijing, China, 11 - 13 June, 2002.

Meehl, G.A., Zwiers, F., Evans, J., Knutson, T. et al. (2000) Trends in extreme weather and climate events: Issues related to modelling extremes in projections of future climate change. *Bull. Amer. Met. Soc*. 81(3), 427-437.

North, G. R. (1984) Empirical orthogonal functions and normal modes. *Journal of the Atmospheric Sciences*, 41, 879-887.

Sampson, P. D., and Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc*. 87, 108–119.

Stainforth, D. A., T. Aina, C. Christensen, M. Collins, N. Faull, D. J. Frame, J. A. Kettleborough, S. Knight, A. Martin, J. M. Murphy, C. Piani, D. Sexton, L. A. Smith, R. A. Spicer, A. J. Thorpe and M. R. Allen (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, 433, 403-406.